ARL-TR-7836 • Sep 2016

US Army Research Laboratory

Constrained Fisher Scoring for a Mixture of Factor Analyzers

by Gene T Whipps, Emre Ertin, and Randolph L Moses

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7836 ● Sep 2016

US Army Research Laboratory

Constrained Fisher Scoring for a Mixture of Factor Analyzers

by Gene T Whipps
Sensors and Electron Devices Directorate, ARL

Emre Ertin and Randolph L Moses
Electrical and Computer Engineering Department, The Ohio State University

REPORT DOCUMENTATION PAGE	***Form Approved*** ***OMB No. 0704-0188***

1. REPORT DATE *(DD-MM-YYYY)* September 2016	**2. REPORT TYPE** Technical Report	**3. DATES COVERED (From - To)** January to September 2016

4. TITLE AND SUBTITLE Constrained Fisher Scoring for a Mixture of Factor Analyzers	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Gene T Whipps, Emre Ertin, and Randolph L Moses	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory Sensors and Electron Devices Directorate ATTN: RDRL-SES-A Adelphi, MD 20783	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-7836
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

primary author's email: <gene.t.whipps.civ@mail.mil>.

14. ABSTRACT

This report considers the problem of learning an object appearance manifold using a spatially distributed network of sensors. Sensor nodes observe an object from different aspects and then learn a joint statistical model for the object manifold. We employ a mixture of factor analyzers model and derive a Fisher scoring method for maximum-likelihood estimation of the model parameters. We analyze convergence of the scoring method and derive stopping conditions for exiting the iterative algorithm. Simulation examples demonstrate that the proposed approach provides faster model learning over the popular expectation-maximization algorithm with similar computational requirements. Lastly, we demonstrate the efficacy of the proposed method for learning a global appearance model across the entire sensor network.

15. SUBJECT TERMS

constrained maximum likelihood estimation, mixture of factor analyzers, Newton's method, expectation maximization

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Gene T Whipps
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	44	**19b. TELEPHONE NUMBER (Include area code)** 301 394 2372

Contents

List of Figures

List of Tables

Acknowledgments

The research reported here was partially supported by a grant from Army Research Office, W911NF-11-1-0391.

1. Introduction

In this effort, we consider the problem of learning low-dimensional representations of objects from a spatially distributed network of sensors. Such a sensor network can be used to construct rich appearance models for objects in their common field of view.[1] These models can then be used in a variety of ways, for example, to identify previously seen objects if they reappear in the network at a later time, or distill important common or discriminating characteristics associated with different objects.

As an example, consider a network of cameras capturing images of an object from different but possibly overlapping aspects, say, as the object traverses through the network's field of view. The ensemble of images captured by the network may be well modeled by a low-dimensional nonlinear manifold in the high-dimensional ambient space of images. One approach to estimating this appearance model might be to learn independent models of a local object manifold at each sensor node, and then later share these models across the network. Such an approach is likely to produce models with high uncertainty or even gaps if a given sensor node observes the object for only a limited set of aspects. An alternative approach, and the one we pursue in this report, is to construct a single joint model for the image ensemble across the network. The parameter estimates of the joint model will improve with the number of sensor nodes,[2] since the number of unknown parameters in the model is intrinsic to the object and fixed, whereas the measurements scale linearly with the number of sensor nodes.

We model the overall statistics of the observations, as seen across multiple aspects and multiple sensors, as a mixture of factor analyzers (MFA)[3] and derive a centralized gradient-based algorithm for learning model parameters. The MFA model is both probabilistic and generative, and can be used for dimensionality reduction, manifold learning, and signal recovery from compressed sensing.[1] In the case of learning a data manifold, the MFA model is a linearization of a potentially nonlinear structure. Each MFA factor mean relates to a point on the manifold, while the tangent plane of the manifold at that point is spanned by the columns of that factor's loading matrix.

Factor analysis has been used successfully in many problems.[1,4] Ghahramani and Hinton[3] applied the expectation-maximization (EM) algorithm to parameter estimation for the MFA model. The EM algorithm[5] is a popular iterative method for *maximum-likelihood* (ML) estimation with local convergence properties and a simple implementation for many applications in statistical signal processing. However, in many practical scenarios, it can exhibit slow convergence,[6] which leads to research in acceleration methods, notably *hybrid methods*[7] that complement EM with information from the problem's likelihood and its gradient. In this report, we derive a constrained Fisher scoring method that can be viewed as a hybrid approach since a subset of the parameter updates have an equivalence with the EM algorithm for the MFA model.

The relationship between the method of scoring and the EM algorithm for the exponential family of distributions was first described by Titterington.[8] The specific relationship for a Gaussian mixture model (GMM) was later formulated by Xu and Jordan,[9] where they showed EM steps can be related to the score function in the parameter space through a positive definite scaling matrix. In Xu and Jordan,[9] the scaling matrix was teased out from the EM update equations. Alternatively, one can show that this matrix is the expected Fisher information matrix (FIM) of the *complete* data model for a GMM. In this case, the scoring method provides a more procedural approach for determining the gradient-scaling matrix and enables convergence analysis using standard tools from optimization theory.

The contributions of this report are as follows.

First, we develop a computationally attractive method of scoring for estimating the parameters of a MFA model from measurements collected by a spatially distributed sensor network. The algorithm is a *centralized* approach in that it assumes that the sensed data can be accumulated at a single, central point for joint data processing. The proposed scoring algorithm is derived using the complete data FIM, as opposed to the incomplete data FIM. The complete data FIM has a block-diagonal structure, leading to significantly reduced computations compared to Newton's method. Furthermore, the FIM exposes a low-rank structure that permits further reductions in computations. The scoring method is shown to have faster convergence than the EM algorithm for a MFA,[3] sometimes significantly faster, while retaining comparable computational complexity.

Second, we demonstrate the efficacy of the constrained scoring approach for efficiently federating across the entire sensor network a global appearance model of objects, even if each sensor observes only a very limited subset of the entire model appearance. Thus, the approach presented in this report is an efficient method for learning a global model. Previously developed methods[10] provided learning of aspect-independent object signatures; in this work, we develop a method for learning appearance models with aspect dependence.

The remainder of this report is outlined as follows. In Section 2, we outline the MFA observation model. In Section 3, we derive the update equations of the centralized scoring method for the MFA model. In Section 4, numerical examples demonstrate the improved performance of the centralized algorithm over EM. Finally, conclusions are given in Section 5.

2. Mixture of Factor Analyzers

Consider a set of M spatially distributed sensors that observe a scene. Sensor nodes collect N vector observations each, denoted $\boldsymbol{x}_{mi} \in \mathbb{R}^p$, for $m = 1, 2, \ldots, M$ and $i = 1, 2, \ldots, N$. The observations are statistically independent, but not identically distributed across the sensor network. Each observation from the m^{th} sensor node is modeled as a mixture of Gaussians, with likelihood given by

$$p_m(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_{mk} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_k, \Sigma_k), \tag{1}$$

where

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left(-\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}} \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) \right). \tag{2}$$

The covariance takes on a low-rank structure according to

$$\Sigma_k = \psi_k \mathrm{I}_p + \Lambda_k \Lambda_k^{\mathrm{T}}, \tag{3}$$

where $\psi_k > 0$ and $\Lambda_k \in \mathbb{R}^{p \times r}$, with $r < p$. The matrix Λ_k is called the *factor loading* matrix and ψ_k is called the *uniqueness*.[11] The factor loading matrix characterizes the lower-dimensional latent space and the uniquenesses account for observation noise and imperfect modeling as a low-rank structure. Each α_{mk} represents the mixing proportion of factor k that sensor m observes. Each sensor node observes the common factors with potentially differing proportions.

The vector $\boldsymbol{\theta}$ consists of the parameters $\boldsymbol{\alpha}_m = [\alpha_{m1}, \ldots, \alpha_{mK}]^{\mathrm{T}}$ for $m = 1, 2, \ldots, M$ and $\boldsymbol{\xi}_k = \left[\boldsymbol{\mu}_k^{\mathrm{T}}, \boldsymbol{\lambda}_k^{\mathrm{T}}, \psi_k\right]^{\mathrm{T}}$ for $k = 1, \ldots, K$ where $\boldsymbol{\lambda}_k = \mathrm{vec}\,(\Lambda_k)$. We specify the vector of unknown parameters as

$$\boldsymbol{\theta} = \left[\boldsymbol{\xi}_1^{\mathrm{T}}, \boldsymbol{\xi}_2^{\mathrm{T}}, \ldots, \boldsymbol{\xi}_K^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_M^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{4}$$

The *log-likelihood* function of $\boldsymbol{\theta}$ given observation $\boldsymbol{x}$ at sensor node m is given by

$$\ell_m(\boldsymbol{\theta}; \boldsymbol{x}) = \log p_m(\boldsymbol{x}; \boldsymbol{\theta}). \tag{5}$$

Given N independent observations from each of the M sensors, the data log-likelihood function is then

$$\ell(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \ell_m(\boldsymbol{\theta}; \boldsymbol{x}_{mi}), \tag{6}$$

where $\boldsymbol{X}$ represents the collection of NM observations from across the sensor network.

To simplify the notation, it is assumed the sensor nodes all collect the same number of observations N. However, the model and the proposed algorithms that follow can be readily modified to accommodate differing numbers of observations.

It is common to treat the MFA model as having latent variables, variables that are not directly observed. The MFA model can be interpreted as having both continuous and discrete latent variables (e.g., see McLachlan and Peel[12]). We consider only the discrete latent terms. We denote by $z \in \{1, 2, \ldots, K\}$ the index to which factor generated observation $\boldsymbol{x}$. Instead of modeling continuous latent terms to describe the factors, the low-rank structure in Eq. 3 is favored as a parametric model of each factor's covariance. The joint distribution of the complete data pair $\{\boldsymbol{x}, z\}$ at sensor node m is given by

$$p_m(\boldsymbol{x}, z; \boldsymbol{\theta}) = p_m(z; \boldsymbol{\theta}) p(\boldsymbol{x}|z; \boldsymbol{\theta}) = \alpha_{mz} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_z, \Sigma_z). \tag{7}$$

The *complete* data log-likelihood of $\boldsymbol{\theta}$, given the MN independent observation pairs, is given by

$$\ell^c(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \log \alpha_{mz_{mi}} \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_{z_{mi}}, \Sigma_{z_{mi}}). \tag{8}$$

3. Constrained Fisher Scoring

It is well known that no closed-form solution exists for ML estimates of the parameters of an MFA. Instead, we must rely on iterative algorithms, such as Newton's method or EM. While Newton's method may converge to a solution quickly, the computational complexity tends to be impractical. On the other hand, the computational complexity of EM per iteration tends to be favorable, but its convergence rate can be less favorable. As a balance between Newton's method and EM, we consider Fisher's method of scoring. Similar to EM but unlike Newton's method, the scoring method requires only first-order derivatives. Though like Newton's method, the performance of the scoring method, in terms of convergence and convergence rate, can be analyzed using standard techniques from optimization theory.

As an unconstrained problem, parameters of the MFA are unidentifiable.[11] Therefore, we must impose constraints. One immediate constraint is the mixing proportions be proper probabilities (i.e., $\alpha_{mk} \geq 0$ and $\sum_k \alpha_{mk} = 1$). Additionally, the uniquenesses must be positive and the factor loading matrix must be restricted to achieve identifiability. It is easy to show that there are an infinity of solutions for each factor loading matrix Λ with equivalent product $\Lambda\Lambda^{\mathrm{T}}$. There are many ways to constrain the factor loading matrix to ensure identifiability.[13] One particularly useful condition is that the factor loading matrix be a lower-triangular matrix, which removes the indeterminacy due to matrix rotations.[13]

We denote by $\boldsymbol{f}(\boldsymbol{\theta})$ the vector of constraints on the unknown parameters such that $\boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}$. The constrained ML problem is given by

$$\max_{\boldsymbol{\theta}} \ \ell(\boldsymbol{\theta}; \boldsymbol{X}) \quad \text{s.t.} \quad \boldsymbol{f}(\boldsymbol{\theta}) = \boldsymbol{0}. \tag{9}$$

The maximizer of Eq. 9 can be solved iteratively via the constrained scoring method.[14]

The constrained scoring method is an iterative, gradient-based algorithm given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left(U_{\boldsymbol{\theta}} \left(U_{\boldsymbol{\theta}}^{\mathrm{T}} J_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}} \right)^{-1} U_{\boldsymbol{\theta}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}, \tag{10}$$

where $J_{\boldsymbol{\theta}}$ is the (unconstrained) expected FIM, defined by

$$J_{\boldsymbol{\theta}} = \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \right), \tag{11}$$

matrix $U_{\boldsymbol{\theta}}$ is defined by the constraints on $\boldsymbol{\theta}$, and $\nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{X})$ is the score function of the (incomplete) data log-likelihood. In Moore et al.[14] and Stoica and Ng,[15] the matrix $U_{\boldsymbol{\theta}}$ is required to be orthonormal (i.e., $U_{\boldsymbol{\theta}}^{\mathrm{T}} U_{\boldsymbol{\theta}} = I$). However, as shown in Appendix A, the orthonormality requirement is unnecessary.

Instead of using the FIM in Eq. 11, we propose substituting the complete data information matrix. The complete data FIM is defined according to

$$J_{\boldsymbol{\theta}}^{c} = \mathrm{E}\left(\nabla_{\boldsymbol{\theta}} \ell^{c}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \ell^{c}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) \right). \tag{12}$$

This substitution was previously proposed for recursive estimation with incomplete data.[8] Though not proven directly, Titterington states equivalence between batch-mode EM and (unconstrained) Fisher scoring when substituting the complete data FIM and when the complete data can be expressed as an exponential family distribution.[8] We do not arrive at identical update equations between unconstrained scoring and EM. Instead, we find equivalence for a subset of the equations between EM and *constrained* scoring. In particular, the iterates for the mixing proportions are identical between EM and constrained scoring. The iterates for the each factor's means are equivalent in an asymptotic sense,

In Eqs. 11 and 12, the expectations are with respect to $\boldsymbol{X}$ and $(\boldsymbol{X}, \boldsymbol{Z})$, respectively. For data set $\boldsymbol{X}$, the factor indicators are not observed directly. In this view, the FIM in Eq. 11 is of the *incomplete* data, whereas Eq. 12 is of the *complete* data. For the remainder of this report, references to the FIM imply the expected information matrix of the complete data in Eq. 12. The proposed scoring method is then given by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left(U_{\boldsymbol{\theta}} \left(U_{\boldsymbol{\theta}}^{\mathrm{T}} J_{\boldsymbol{\theta}}^{c} U_{\boldsymbol{\theta}} \right)^{-1} U_{\boldsymbol{\theta}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}. \tag{13}$$

For the MFA model, the FIM $J^c_{\boldsymbol{\theta}}$ has a closed-form expression and can be determined through a change of variables from the standard GMM. Define $\boldsymbol{\Phi} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\sigma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{T}}, \boldsymbol{\sigma}_K^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_M^{\mathrm{T}}]^{\mathrm{T}}$, where $\boldsymbol{\sigma}_k = \mathrm{vec}\,(\Sigma_k)$. The information matrix of the MFA model is given by

$$J^c_{\boldsymbol{\theta}} = G^{\mathrm{T}}_{\boldsymbol{\theta}} J^c_{\boldsymbol{\Phi}} G_{\boldsymbol{\theta}}, \tag{14}$$

where $J^c_{\boldsymbol{\Phi}}$ is the complete data FIM of the GMM defined by

$$J^c_{\boldsymbol{\Phi}} = \mathrm{E}\left(\nabla_{\boldsymbol{\Phi}} \ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z}) \nabla^{\mathrm{T}}_{\boldsymbol{\Phi}} \ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z})\right), \tag{15}$$

and $G_{\boldsymbol{\theta}}$ is the Jacobian matrix for the change of variables defined by

$$G_{\boldsymbol{\theta}} = \frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}. \tag{16}$$

We admit a slight abuse of notation in Eq. 15 by substituting $\boldsymbol{\Phi}$ for $\boldsymbol{\theta}$ in ℓ^c from Eq. 8. This is for notational convenience intended to simply imply the complete data log-likelihood function of the GMM without the low-rank structure in Eq. 3 imposed on the covariance.

The details of $J^c_{\boldsymbol{\Phi}}$ and $G_{\boldsymbol{\theta}}$ are provided in Appendix B, where $J^c_{\boldsymbol{\Phi}}$, the FIM of the GMM, is shown to be block diagonal. It is clear from the definitions of $\boldsymbol{\theta}$ and $\boldsymbol{\Phi}$ that a change of variables is necessary only for each factor's covariance. Thus, the FIM of the MFA has the block-diagonal form

$$J^c_{\boldsymbol{\theta}} = \mathrm{diag}\left(J_{\boldsymbol{\mu}_1}, (G^{\mathrm{T}}_{\boldsymbol{\sigma}_1} J_{\boldsymbol{\sigma}_1} G_{\boldsymbol{\sigma}_1}), \ldots, J_{\boldsymbol{\mu}_K}, (G^{\mathrm{T}}_{\boldsymbol{\sigma}_K} J_{\boldsymbol{\sigma}_K} G_{\boldsymbol{\sigma}_K}), J_{\boldsymbol{\alpha}_1}, \ldots, J_{\boldsymbol{\alpha}_M}\right). \tag{17}$$

This represents the unconstrained information matrix of the complete data model. However, the MFA is overparameterized and not identifiable.[11,13] While the factor means may remain unconstrained, the factor loading matrices, the uniquenesses, and the mixing proportions should be constrained. The details of these constraints are provided in the following subsections. First, we remark that the constraints are decoupled across the parameters such that the constrained FIM for the MFA takes on the block-diagonal form given by

$$U_{\boldsymbol{\theta}}^{\mathrm{T}} J_{\boldsymbol{\theta}}^{c} U_{\boldsymbol{\theta}} = \mathrm{diag}\Big(J_{\boldsymbol{\mu}_1}, (U_{\boldsymbol{\sigma}_1}^{\mathrm{T}} G_{\boldsymbol{\sigma}_1}^{\mathrm{T}} J_{\boldsymbol{\sigma}_1} G_{\boldsymbol{\sigma}_1} U_{\boldsymbol{\sigma}_1}), J_{\boldsymbol{\mu}_2}, (U_{\boldsymbol{\sigma}_2}^{\mathrm{T}} G_{\boldsymbol{\sigma}_2}^{\mathrm{T}} J_{\boldsymbol{\sigma}_2} G_{\boldsymbol{\sigma}_2} U_{\boldsymbol{\sigma}_2}), \dots, \\ J_{\boldsymbol{\mu}_K}, (U_{\boldsymbol{\sigma}_K}^{\mathrm{T}} G_{\boldsymbol{\sigma}_K}^{\mathrm{T}} J_{\boldsymbol{\sigma}_K} G_{\boldsymbol{\sigma}_K} U_{\boldsymbol{\sigma}_K}), (U_{\boldsymbol{\alpha}_1}^{\mathrm{T}} J_{\boldsymbol{\alpha}_1} U_{\boldsymbol{\alpha}_1}), \dots, (U_{\boldsymbol{\alpha}_M}^{\mathrm{T}} J_{\boldsymbol{\alpha}_M} U_{\boldsymbol{\alpha}_M}) \Big) . \tag{18}$$

As a result of this form, the parameter iterates of a factor are decoupled from those of other factors. Also, for a given factor, the iterates of the means and mixing probability decouple. In contrast, the parameters of the structured covariance remain coupled. Subsequently, the constrained scoring method in Eq. 13 factors into a set of update equations given by

$$\boldsymbol{\alpha}_m^{(t+1)} = \boldsymbol{\alpha}_m^{(t)} + \Big(U_{\boldsymbol{\alpha}_m} \left(U_{\boldsymbol{\alpha}_m}^{\mathrm{T}} J_{\boldsymbol{\alpha}_m} U_{\boldsymbol{\alpha}_m} \right)^{-1} U_{\boldsymbol{\alpha}_m}^{\mathrm{T}} \nabla_{\boldsymbol{\alpha}_m} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} , \tag{19}$$

$$\boldsymbol{\mu}_k^{(t+1)} = \boldsymbol{\mu}_k^{(t)} + \Big(J_{\boldsymbol{\mu}_k}^{-1} \nabla_{\boldsymbol{\mu}_k} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} , \tag{20}$$

$$\begin{bmatrix} \boldsymbol{\lambda}_k^{(t+1)} \\ \psi_k^{(t+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_k^{(t)} \\ \psi_k^{(t)} \end{bmatrix} + \Big(U_{\boldsymbol{\sigma}_k} \left(U_{\boldsymbol{\sigma}_k}^{\mathrm{T}} G_{\boldsymbol{\sigma}_k}^{\mathrm{T}} J_{\boldsymbol{\sigma}_k} G_{\boldsymbol{\sigma}_k} U_{\boldsymbol{\sigma}_k} \right)^{-1} U_{\boldsymbol{\sigma}_k}^{\mathrm{T}} G_{\boldsymbol{\sigma}_k}^{\mathrm{T}} \nabla_{\boldsymbol{\sigma}_k} \ell(\boldsymbol{\Phi}; \boldsymbol{X}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} , \tag{21}$$

where $\boldsymbol{\sigma}_k^{(t)}$ is evaluated by substituting $(\boldsymbol{\lambda}_k^{(t)}, \psi_k^{(t)})$ in Eq. 3. Though they appear fairly complex, the update equations significantly simplify. The details of the update equations and their simplification are provided in the following subsections.

3.1 Mixing Proportions

In this section, we derive an explicit expression for iterates of the mixing proportions that is simplified compared to Eq. 19 and achieves the desired probability constraints. The iterates of each mixing probability basically reduce to sample averages of posterior probabilities of an observation belonging to a corresponding factor.

We briefly drop the dependence on the specific sensor node m to simplify the notation and note that the following results are identical for each $m = 1, \dots, M$. We define $\boldsymbol{\alpha} = [\alpha_1, \dots, \alpha_K]^{\mathrm{T}}$ as the vector of mixing proportions and $\tilde{\boldsymbol{\alpha}} = [\alpha_1, \dots, \alpha_{K-1}]^{\mathrm{T}}$ as the reduced parameter vector.

The probability constraint can be applied through a change of variables since we have a parametric form of $\boldsymbol{\alpha}$ with respect to the reduced parameter vector $\tilde{\boldsymbol{\alpha}}$.[16]

From the sum-to-one condition, we have $\alpha_K = 1 - \mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}$, and for the change of variables we have

$$U_{\boldsymbol{\alpha}} = \frac{\partial \left[\tilde{\alpha}^{\mathrm{T}}, 1 - \mathbf{1}^{\mathrm{T}}\tilde{\boldsymbol{\alpha}}\right]^{\mathrm{T}}}{\partial \tilde{\alpha}^{\mathrm{T}}} = \begin{bmatrix} \mathrm{I}_{K-1} \\ -\mathbf{1}^{\mathrm{T}} \end{bmatrix}. \tag{22}$$

From Eq. B-8 in Appendix B, the term $J_{\boldsymbol{\alpha}}$ for the mixing proportions from the FIM is given by

$$J_{\boldsymbol{\alpha}} = \operatorname{diag}\left(\boldsymbol{\alpha}\right)^{-1}. \tag{23}$$

From Eqs. 22 and 23, the inverse of matrix $U_{\boldsymbol{\alpha}}^{\mathrm{T}} J_{\boldsymbol{\alpha}} U_{\boldsymbol{\alpha}}$ evaluates to

$$\begin{aligned}\left(U_{\boldsymbol{\alpha}}^{\mathrm{T}} J_{\boldsymbol{\alpha}} U_{\boldsymbol{\alpha}}\right)^{-1} &= \frac{1}{N}\left(\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right)^{-1} + \alpha_K^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)^{-1} \\ &= \frac{1}{N}\Big(\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right) - \operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right)\mathbf{1}\left(\alpha_K + \mathbf{1}^{\mathrm{T}}\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right)\mathbf{1}\right)^{-1}\mathbf{1}^{\mathrm{T}}\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right)\Big) \end{aligned} \tag{24}$$

$$= \frac{1}{N}\left(\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}}\right). \tag{25}$$

Equation 24 follows from application of the Woodbury matrix identity and Eq. 25 follows since $\alpha_K + \mathbf{1}^{\mathrm{T}}\operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right)\mathbf{1} = \sum_{k=1}^{K}\alpha_K = 1$. Lastly, it is straightforward to show that

$$\begin{aligned}U_{\boldsymbol{\alpha}}\left(U_{\boldsymbol{\alpha}}^{\mathrm{T}} J_{\boldsymbol{\alpha}} U_{\boldsymbol{\alpha}}\right)^{-1} U_{\boldsymbol{\alpha}}^{\mathrm{T}} &= \frac{1}{N}\begin{bmatrix} \operatorname{diag}\left(\tilde{\boldsymbol{\alpha}}\right) - \tilde{\boldsymbol{\alpha}}\tilde{\boldsymbol{\alpha}}^{\mathrm{T}} & -\alpha_K\tilde{\boldsymbol{\alpha}} \\ -\alpha_K\tilde{\boldsymbol{\alpha}}^{\mathrm{T}} & \alpha_K - \alpha_K^2 \end{bmatrix} \\ &= \frac{1}{N}\left(\operatorname{diag}\left(\boldsymbol{\alpha}\right) - \boldsymbol{\alpha}\boldsymbol{\alpha}^{\mathrm{T}}\right). \end{aligned} \tag{26}$$

In Xu and Jordan,[9] it was shown that Eq. 26 is positive definite provided $\boldsymbol{\alpha}$ is constrained to a probability simplex. Furthermore, it is shown by Xu and Jordan[9] that the iteration in Eq. 19 with the scaling matrix in Eq. 26 simplifies to

$$\alpha_{mk}^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} w_{mik}^{(t)}, \tag{27}$$

where

$$w_{mik} = \frac{\alpha_{mk}\mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K}\alpha_{mj}\mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_j, \Sigma_j)}, \tag{28}$$

for $k = 1, \ldots, K$, $m = 1, \ldots, M$, and $w_{mik}^{(t)}$ is Eq. 28 evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. From Eq. 28, it is easy to see that Eq. 27 and therefore Eq. 19 meet the probability constraints. Subsequently, Eq. 26 is positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Thus, the iteration in Eq. 19 strictly increases the likelihood with respect to the mixing proportions, at least locally.

3.2 Factor Means

Similar to the mixing proportions, the update equations for the factor means simplify. The score function with respect to $\boldsymbol{\mu}_k$ equates to

$$\nabla_{\boldsymbol{\mu}_k} \ell(\boldsymbol{\theta}; \boldsymbol{X}) = \sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik} \Sigma_k^{-1} (\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k), \tag{29}$$

for $k = 1, \ldots, K$. From Eq. B-7 in Appendix B, the term $J_{\boldsymbol{\mu}_k}$ for the factor mean from the FIM is given by

$$J_{\boldsymbol{\mu}_k} = N \sum_{m=1}^{M} \alpha_{mk} \Sigma_k^{-1}. \tag{30}$$

Inserting Eqs. 29 and 30 in Eq. 20, the iteration for the k^{th} factor's mean reduces to

$$\boldsymbol{\mu}_k^{(t+1)} = \boldsymbol{\mu}_k^{(t)} + \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)} (\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k^{(t)})}{N \sum_{m=1}^{M} \alpha_{mk}^{(t)}}. \tag{31}$$

Provided the uniquenesses are positive, the matrix $J_{\boldsymbol{\mu}_k}$ is guaranteed to be positive definite by the definition of the covariance in Eq. 3 when evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. Therefore, the iterations in Eqs. 20 and 31 strictly increase the likelihood with respect to each factor's mean.

3.3 Factor Loading Matrix and Uniqueness

In this section, we focus on detailing the scoring function and the constrained FIM in Eq. 21 with respect to the covariance parameters $(\boldsymbol{\lambda}_k, \psi_k)$. The score function with respect to the parameters $(\boldsymbol{\lambda}_k, \psi_k)$ is found through a change of variables from the score function with respect to $\boldsymbol{\sigma}_k$. It is straightforward to show that the score function with respect to $\boldsymbol{\sigma}_k$ is given by

$$\nabla_{\boldsymbol{\sigma}_k}\ell(\boldsymbol{\Phi};\boldsymbol{X}) = \frac{1}{2}\sum_{m=1}^{M}\sum_{i=1}^{N} w_{mik}\mathrm{vec}\left(\Sigma_k^{-1}(\boldsymbol{x}_{mi}-\boldsymbol{\mu}_k)(\boldsymbol{x}_{mi}-\boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}-\Sigma_k^{-1}\right). \tag{32}$$

We briefly drop the dependence on the specific factor k to simplify the notation and note that the following results are identical for each $k = 1,\ldots,K$. For the change of variables, we have

$$G_{\boldsymbol{\sigma}} = \left[G_{\boldsymbol{\lambda}}, \boldsymbol{g}_{\psi}\right] = \left[\frac{\partial\boldsymbol{\sigma}}{\partial\boldsymbol{\lambda}^{\mathrm{T}}}, \frac{\partial\boldsymbol{\sigma}}{\partial\psi}\right]. \tag{33}$$

The partial derivatives in Eq. 33 evaluate to

$$G_{\boldsymbol{\lambda}} = (\Lambda\otimes\mathrm{I}_p) + (\mathrm{I}_p\otimes\Lambda)\,E, \tag{34}$$

$$\boldsymbol{g}_{\psi} = \left[(\boldsymbol{e}_1^p\otimes\boldsymbol{e}_1^p),(\boldsymbol{e}_2^p\otimes\boldsymbol{e}_2^p),\ldots,\left(\boldsymbol{e}_p^p\otimes\boldsymbol{e}_p^p\right)\right]\mathbf{1}, \tag{35}$$

where $(A\otimes B)$ represents the Kronecker product between matrices A and B and $\boldsymbol{e}_i^a$ is a $(a\times 1)$ unit vector such that all the entries are 0 except the i^{th} entry is 1. The matrix $E\in\mathbb{R}^{pr\times pr}$ is defined by

$$E = \frac{\partial\mathrm{vec}\left(\Lambda^{\mathrm{T}}\right)}{\partial\mathrm{vec}\left(\Lambda\right)^{\mathrm{T}}} = \left[(\mathrm{I}_p\otimes\boldsymbol{e}_1^r),(\mathrm{I}_p\otimes\boldsymbol{e}_2^r),\ldots,(\mathrm{I}_p\otimes\boldsymbol{e}_r^r)\right]. \tag{36}$$

Matrix E is a permutation matrix that satisfies $E^{\mathrm{T}}E = EE^{\mathrm{T}} = \mathrm{I}_{pr}$. Vector $\boldsymbol{g}_{\psi}\in\mathbb{R}^{p^2\times 1}$ simply sums the diagonal elements of a vectorized $(p\times p)$ matrix.

From Eqs. 32 and 33, it is straightforward to show that the score function with respect to the parameters $(\boldsymbol{\lambda},\psi)$ is given by

$$G_{\boldsymbol{\sigma}}^{\mathrm{T}}\nabla_{\boldsymbol{\sigma}}\ell(\boldsymbol{\Phi};\boldsymbol{X}) = \begin{bmatrix} 2\left(\Lambda^{\mathrm{T}}\otimes\mathrm{I}_p\right)\nabla_{\boldsymbol{\sigma}}\ell(\boldsymbol{\Phi};\boldsymbol{X}) \\ \boldsymbol{g}_{\psi}^{\mathrm{T}}\nabla_{\boldsymbol{\sigma}}\ell(\boldsymbol{\Phi};\boldsymbol{X}) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{\lambda}}\ell(\boldsymbol{\theta};\boldsymbol{X}) \\ \nabla_{\psi}\ell(\boldsymbol{\theta};\boldsymbol{X}) \end{bmatrix}. \tag{37}$$

We constrain each factor loading matrix to be lower triangular. This constraint permits *local* identifiability of the factor loading matrices.[13] The lower-triangular constraint is also convenient since it is a linear transformation of the loading matrix in vector form. The reduced parameter vector of the factor loading matrix is defined by

$$\tilde{\boldsymbol{\lambda}} = \left[\tilde{\lambda}_1^{\mathrm{T}}, \tilde{\lambda}_2^{\mathrm{T}}, \ldots, \tilde{\lambda}_r^{\mathrm{T}}\right]^{\mathrm{T}}, \tag{38}$$

where $\tilde{\lambda}_i = [\Lambda_{1,i}, \Lambda_{1,i+1}, \ldots, \Lambda_{1,p}]^{\mathrm{T}}$ is the i^{th} column vector from Λ not including the elements above the diagonal. The constrained factor loading may be expressed according to

$$\boldsymbol{\lambda} = U_{\boldsymbol{\lambda}} \tilde{\boldsymbol{\lambda}}, \tag{39}$$

where $U_{\boldsymbol{\lambda}} \in \mathbb{R}^{pr \times pr - r(r-1)/2}$ simply expands $\tilde{\boldsymbol{\lambda}}$ to $\boldsymbol{\lambda}$ by inserting zeros at the appropriate entries. Specifically, $U_{\boldsymbol{\lambda}}$ is given by

$$U_{\boldsymbol{\lambda}} = \left[\left(\boldsymbol{e}_1^r \otimes \mathcal{I}_1\right), \left(\boldsymbol{e}_2^r \otimes \mathcal{I}_2\right), \ldots, \left(\boldsymbol{e}_r^r \otimes \mathcal{I}_r\right)\right], \tag{40}$$

where

$$\mathcal{I}_i = \begin{bmatrix} \mathbf{0}_{(i-1) \times p-(i-1)} \\ \mathrm{I}_{p-(i-1)} \end{bmatrix}. \tag{41}$$

The term $\mathbf{0}_{a \times b}$ represents a $(a \times b)$ matrix of zeros. It is clear that $U_{\boldsymbol{\lambda}}^{\mathrm{T}} U_{\boldsymbol{\lambda}} = \mathrm{I}_{pr-r(r-1)/2}$ so that $U_{\boldsymbol{\lambda}}^{\mathrm{T}} \boldsymbol{\lambda} = \tilde{\boldsymbol{\lambda}}$.

We do not constrain the uniquenesses to be positive within the gradient steps. Instead, each uniqueness $\psi^{(t+1)}$ is projected onto the interval $[\epsilon, \infty)$ for a small $\epsilon > 0$ after each iteration. As a result, the constraint matrix $U_{\boldsymbol{\sigma}}$ for the gradient update in Eq. 21 is given by

$$U_{\boldsymbol{\sigma}} = \begin{bmatrix} U_{\boldsymbol{\lambda}} & \mathbf{0} \\ \mathbf{0}^{\mathrm{T}} & 1 \end{bmatrix}. \tag{42}$$

In the reduced parameter space, we have

$$U_{\boldsymbol{\sigma}}^{\mathrm{T}} G_{\boldsymbol{\sigma}}^{\mathrm{T}} \nabla_{\boldsymbol{\sigma}} \ell(\boldsymbol{\Phi}; \boldsymbol{X}) = \begin{bmatrix} \nabla_{\tilde{\boldsymbol{\lambda}}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \\ \nabla_{\psi} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \end{bmatrix}. \tag{43}$$

The constrained FIM with respect to a factor's covariance parameters is given by

$$U_{\boldsymbol{\sigma}}^{\mathrm{T}} G_{\boldsymbol{\sigma}}^{\mathrm{T}} J_{\boldsymbol{\sigma}} G_{\boldsymbol{\sigma}} U_{\boldsymbol{\sigma}} = \frac{N}{2} \sum_{m=1}^{M} \alpha_m A^{\mathrm{T}} A, \tag{44}$$

where

$$A = \left(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \right) G_{\boldsymbol{\sigma}} U_{\boldsymbol{\sigma}}. \tag{45}$$

Since $\left(\Sigma^{-1} \otimes \Sigma^{-1} \right)$ is positive definite, a (positive) matrix square root is defined.

Provided Eq. 44 is positive definite, the iteration in Eq. 21 will strictly increase the data likelihood with respect to the covariance parameters. Matrix $A^{\mathrm{T}} A$ is clearly either positive definite or positive *semi*definite. Either way, the iteration in Eq. 21 will not decrease the likelihood. Matrix $A^{\mathrm{T}} A$ is positive definite if A has full column rank. Defining $\left(\Sigma^{-1/2} \otimes \Sigma^{-1/2} \right)$ as the positive root so that it is positive definite, we have the equivalent condition that $A^{\mathrm{T}} A$ is positive definite if $G_{\boldsymbol{\sigma}} U_{\boldsymbol{\sigma}}$ has full column rank.

Inserting Eqs. 44 and 43 in the iteration in Eq. 21, the iteration for the reduced parameters is given by

$$\left[\begin{array}{c} \tilde{\boldsymbol{\lambda}}_k^{(t+1)} \\ \psi_k^{(t+1)} \end{array} \right] = \left[\begin{array}{c} \tilde{\boldsymbol{\lambda}}_k^{(t)} \\ \psi_k^{(t)} \end{array} \right] + \frac{2}{N \sum_{m=1}^{M} \alpha_{mk}^{(t)}} \left(\left(A_k^{\mathrm{T}} A_k \right)^{-1} \left[\begin{array}{c} \nabla_{\tilde{\boldsymbol{\lambda}}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \\ \nabla_{\psi} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \end{array} \right] \right|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}} . \tag{46}$$

3.4 Convergence and Stopping Condition

We briefly discuss algorithm convergence and provide a stopping condition for the iterative algorithm. We set $P_{\boldsymbol{\theta}} = U_{\boldsymbol{\theta}} \left(U_{\boldsymbol{\theta}}^{\mathrm{T}} J_{\boldsymbol{\theta}}^{c} U_{\boldsymbol{\theta}} \right)^{-1} U_{\boldsymbol{\theta}}^{\mathrm{T}}$. For a sufficient number of iterations, the scoring method converges according to[9]

$$\left| \boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{*} \right| \leq \rho_g \left| \boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{*} \right|, \tag{47}$$

if $\rho_g < 1$ for fixed point $\boldsymbol{\theta}^{*}$, where $\rho_g = \left| \mathrm{I} + P_{\boldsymbol{\theta}^{*}} H_{\boldsymbol{\theta}^{*}} \right|$. The term $H_{\boldsymbol{\theta}}$ is the Hessian matrix of the data log-likelihood defined by

$$H_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \ell(\boldsymbol{\theta}; \boldsymbol{X}). \tag{48}$$

The rate ρ_g depends on the form of matrix $P_{\boldsymbol{\theta}^*}$ and how well it conditions $H_{\boldsymbol{\theta}^*}$. Using the triangle inequality and repeated application of Eq. 47, the change between consecutive iterates is upper-bounded, for large t, according to

$$\left|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}\right| \leq \rho_g^t C_g, \tag{49}$$

where $C_g = (1 + \rho_g) \left|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\right| > 0$. If $\rho_g < 1$, this change-error also predictably contracts. Unlike Eq. 47, the left-hand side of Eq. 49 does not require knowledge of $\boldsymbol{\theta}^*$. Therefore, we rely on the change-error to exit the gradient steps. For a desired tolerance ϵ_g, the algorithm exits when the condition $\left|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}\right| < \epsilon_g$ is met.

3.5 Algorithm Summary

Figure 1 summarizes the constrained Fisher scoring algorithm for an MFA.

```
Input: Each sensor $m$ measures $\boldsymbol{x}_{mi}$ for $i = 1, \ldots, N$
Output: Initialization: $t = 0$, $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\theta}^{(-1)}$ s.t. $|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^{(-1)}| > \epsilon_g$
  while $|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^{(t-1)}| > \epsilon_g$ do
    Update parameters via constrained Fisher scoring
    for $k = 1, \ldots, K$, $m = 1, \ldots, M$, $i = 1, \ldots, N$ do
      Calculate posterior probability of $\boldsymbol{x}_{mi}$ belonging to factor $k$ given $\boldsymbol{\theta}^{(t)}$
      $w_{mik}^{(t)} = \frac{\alpha_{mk}^{(t)} \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \alpha_{mj}^{(t)} \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}$
    end for
    for $k = 1, \ldots, K$ do
      Update $k^{\text{th}}$ factor's parameters
      $\boldsymbol{\mu}_k^{(t+1)} = \boldsymbol{\mu}_k^{(t)} + \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)} (\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k^{(t)})}{N \sum_{m=1}^{M} \alpha_{mk}^{(t)}}$
      $\begin{bmatrix} \tilde{\boldsymbol{\lambda}}_k^{(t+1)} \\ \psi_k^{(t+1)} \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{\lambda}}_k^{(t)} \\ \psi_k^{(t)} \end{bmatrix} + \frac{2}{N \sum_{m=1}^{M} \alpha_{mk}^{(t)}} \Bigg( (A_k^{\mathrm{T}} A_k)^{-1} \begin{bmatrix} \nabla_{\tilde{\boldsymbol{\lambda}}} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \\ \nabla_{\psi} \ell(\boldsymbol{\theta}; \boldsymbol{X}) \end{bmatrix} \Bigg|_{\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}}$
      $\boldsymbol{\lambda}_k^{(t+1)} = U_{\boldsymbol{\lambda}} \tilde{\boldsymbol{\lambda}}_k^{(t+1)}$
      if $\psi_k^{(t+1)} \leq 0$ then
        $\psi_k^{(t+1)} = \epsilon$
      end if
    end for
    for $m = 1, \ldots, M$ do
      Update mixing probabilities for sensor $m$
      $\boldsymbol{\alpha}_m^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} [w_{mi1}^{(t)}, \ldots, w_{miK}^{(t)}]^{\mathrm{T}}$
    end for
    $t \leftarrow t + 1$
  end while
  return $\boldsymbol{\theta}^{(t+1)}$
```

Fig. 1 Constrained Fisher scoring for an MFA

3.6 Relationship with Expectation-Maximization

The constrained scoring iteration for the mixing proportions is identical to that for the EM algorithm.[3] Furthermore, the iteration for the factor means are equivalent to the EM algorithm in an asymptotic sense. If in Eq. 31 we substitute $\sum_{m=1}^{M} \alpha_{mk}^{(t+1)}$ for $\sum_{m=1}^{M} \alpha_{mk}^{(t)}$, and note that

$$N \sum_{m=1}^{M} \alpha_{mk}^{(t+1)} = \sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)}, \tag{50}$$

the iterates for each $k = 1, 2, \ldots, K$ factor mean becomes

$$\begin{aligned} \boldsymbol{\mu}_k^{(t+1)} &= \boldsymbol{\mu}_k^{(t)} + \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)} (\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k^{(t)})}{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)}} \\ &= \frac{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)} \boldsymbol{x}_{mi}}{\sum_{m=1}^{M} \sum_{i=1}^{N} w_{mik}^{(t)}}. \end{aligned} \tag{51}$$

This iteration for the factor mean is exactly the EM-based iteration.[3]

Since $\boldsymbol{\alpha}_m^{(t+1)}$, for each $m = 1, 2, \ldots, M$, converges to a solution, the sequence $\{\boldsymbol{\alpha}_m^{(0)}, \boldsymbol{\alpha}_m^{(1)}, \ldots, \boldsymbol{\alpha}_m^{(t)}\}$ is a Cauchy sequence. Thus, the difference $\boldsymbol{\alpha}_m^{(t+1)} - \boldsymbol{\alpha}_m^{(t)}$ can be made arbitrarily small for a sufficiently large t. Hence, in an asymptotic sense, the factor mean updates in Eqs. 31 and 51 are equivalent.

In terms of each factor's covariance parameters, the constrained scoring and EM methods differ. The EM-based iterates for the factor uniqueness and loading matrix are decoupled,[3] while the iterates from the constrained scoring method are coupled. This difference may also manifest in the difference in convergence rates seen in the simulation examples in the following section.

4. Simulation Examples

We present 2 simulation examples that illustrate the effectiveness of the proposed centralized scoring method. We first consider a synthetic MFA example under different observation noise conditions to compare algorithm convergence and computation. We then consider a multi-aspect observation model that demonstrates the benefits of information sharing in global model learning.

4.1 Synthetic MFA Example

We first compare the performance of the proposed algorithms using the generative MFA model. Parameter estimates are generated using the centralized EM algorithm in Whipps et al.,[17] a constrained Newton's method, and the proposed constrained Fisher scoring method. The constrained Newton's method implements the gradient step according to

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t+1)} + \left(U_{\boldsymbol{\theta}}(-U_{\boldsymbol{\theta}}^{\mathrm{T}} H_{\boldsymbol{\theta}} U_{\boldsymbol{\theta}})^{-1} U_{\boldsymbol{\theta}}^{\mathrm{T}} \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}; \boldsymbol{X})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}, \tag{52}$$

where $H_{\boldsymbol{\theta}}$ is the Hessian matrix of the data log-likelihood defined by

$$H_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \ell(\boldsymbol{\theta}; \boldsymbol{X}). \tag{53}$$

In general terms, the EM algorithm for the MFA model is simple computation-wise, but tends to converge slowly. Newton's method typically converges faster than first-order methods, but at a higher computational cost per iteration. We demonstrate that the constrained Fisher scoring has EM-like computational requirements with improved convergence properties.

We consider a simple MFA model. The sensor observations are 3 dimensional ($p = 3$) with a 2-D latent space ($r = 2$) and 2 mixed factors ($J = 2$). The latent object is composed of 2 planar segments in the shape of a "L" that intersect along the line between points $(-1, 0, 0)$ and $(1, 0, 0)$. Specifically, the parameters of the 2-component mixture are given by

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \tag{54}$$

$$\Lambda_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \Lambda_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \tag{55}$$

$$\psi_1 = \psi_2 = \psi, \tag{56}$$

$$\boldsymbol{\alpha}_m = [0.75 - m0.05,\ 0.25 + m0.05]^{\mathrm{T}}, \tag{57}$$

for $m = 1, 2, \ldots, 9$. A similar simulation example was used by Baek and McLachlan[18] to measure performance for the single-sensor case.

For these simulations, there are $M = 9$ sensor nodes. For each of the $M = 9$ sensor nodes, $N = 50$ samples are generated using the generative MFA model. The algorithms do not attempt to account for the relationship $\boldsymbol{\lambda}_k = \boldsymbol{\mu}_k$ of the simulated mixture.

The algorithms are initialized by the true values perturbed by small, independent errors. The algorithms are initialized with $\boldsymbol{\theta}^{(0)} = \mathcal{P}(\boldsymbol{\theta} + \boldsymbol{\nu}\ )$, where $\mathcal{P}(\boldsymbol{\theta})$ projects $\boldsymbol{\theta}$ onto the feasible set (i.e., mixing proportions are proper probabilities, the uniquenesses are positive, and the factor loading matrices are lower-triangular). $\boldsymbol{\nu}$ is a realization of spherical noise distributed according to $\boldsymbol{\nu} \sim \mathcal{N}(0, \sigma^2 \mathrm{I})$ with $\sigma = 10^{-2}$. We note that iterative algorithms such as EM and gradient ascent are sensitive to their initialization, and ML problems typically have multiple stationary points. The proposed algorithm is no different in this regard. The k-means algorithm can be viewed as a simplification of EM[19] and can provide a fast initialization for all 3 algorithms. However, we do not explore initialization methods here. To make fair comparisons, the initial values are close to the true values and are the same for each algorithm.

Figures 2–4 are results from a single set of $NM = 450$ samples with each factor's uniqueness at $\psi = 1/32$. The sample points are plotted in Fig. 2 as markers (each marker type corresponds to a sensor node) along with the 2 true latent line segments. Figure 2 is a 2-D perspective of the 3-D sensor data. As seen in Fig. 2, the data points are concentrated about their factors save a few points near the vertex.

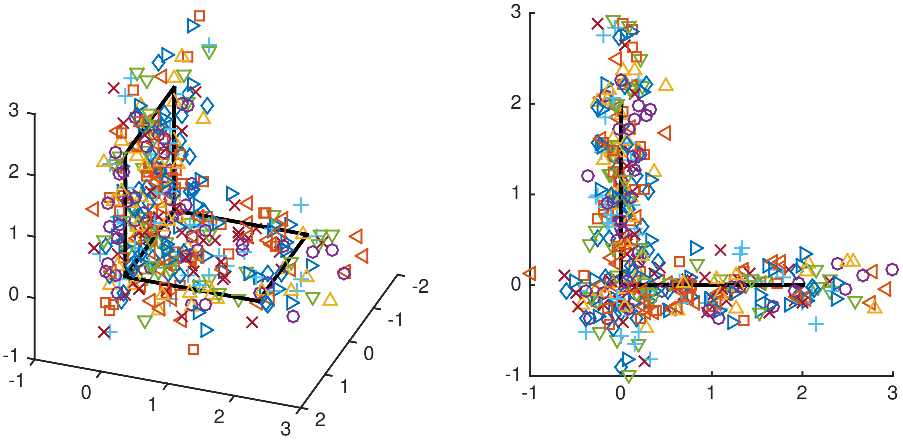

Fig. 2 3-D and 2-D views of simulated sensor data of the simulated MFA with uniqueness $\psi = 1/32$ from $M = 9$ sensor nodes.

We first compare the performance of the proposed scoring algorithm with (constrained) Newton's method and the EM algorithm. The iterative algorithms are stopped after the error $|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}|$ falls below $\epsilon_g = 10^{-12}$.

Figure 3 demonstrates algorithm convergence by showing the error $|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)}|$ versus gradient iterations from the methods of constrained Newton ($+$), EM ($\triangle$), and constrained Fisher scoring ($\circ$). As seen in Fig. 3, the change-error from both Newton's method and Fisher scoring decreases much faster than EM, with Newton's method having the steepest slope.

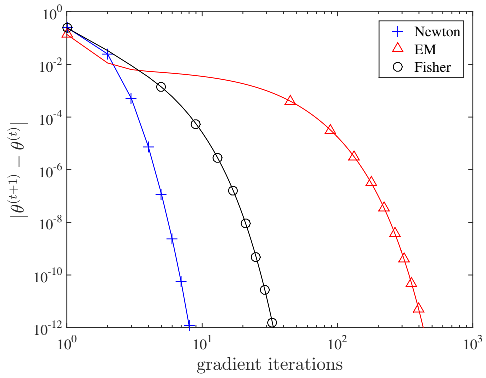

Fig. 3 Change-error vs. gradient iterations from constrained Newton ($+$), EM ($\triangle$), and constrained Fisher ($\circ$) with uniqueness $\psi = 1/32$

Figure 4 plots the execution times in milliseconds reported by MATLAB for each gradient step of each algorithm and illustrates the relative computation times for the algorithms. As seen in Fig. 4, Fisher scoring and EM execute a gradient step in nearly equal time and approximately 10 times faster than Newton's method.

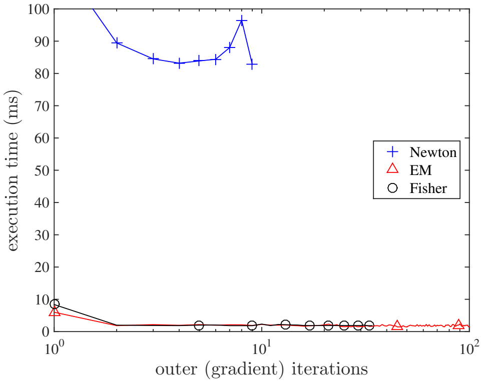

Fig. 4 Execution times per gradient step of constrained Newton ($+$), EM ($\triangle$), and constrained Fisher ($\circ$) with uniqueness $\psi = 1/32$

Table 1 provides total algorithm execution times reported by MATLAB to reach the error tolerance $\epsilon_g = 10^{-12}$, sample averaged over 100 Monte Carlo trials; shown in parenthesis are the execution time standard deviations. The 2 columns of execution times in Table 1 correspond to uniquenesses of $\psi = 1/32$ and $\psi = 1/8$, respectively. As seen in Table 1, Fisher scoring reaches the error tolerance in the shortest amount of execution time of all the methods.

Table 1 Average total execution times in milliseconds to reach a tolerance of $\epsilon_g = 10^{-12}$. The total times are sample-averaged over 100 Monte Carlo trials with the sample deviation listed in parenthesis.

Method	$\boldsymbol{\psi = 1/32}$	$\boldsymbol{\psi = 1/8}$
Newton	1288 (1404)	1551 (1526)
EM	992 (1134)	310 (85)
Fisher	58 (10)	130 (26)

Figures 5–7 and Table 1 show convergence and computation results for the previous example, but here we quadruple the factor uniqueness to $\psi = 1/8$. As seen in Fig. 5, it is more difficult to visually assign many of the samples to a specific factor with $\psi = 1/8$. Thus, we might expect longer convergence times in this case. This is illustrated by comparing the convergence of Fisher scoring between Figs. 3 and 6, where we see somewhat longer convergence times with $\psi = 1/8$. Curiously, the convergence rate of EM appears improved, but remains slower than Fisher scoring. The total execution time of Fisher scoring remains shorter than both Newton's method and EM, as seen in the last column of Table 1 with $\psi = 1/8$. Fisher scoring outperforms EM in terms of convergence rate and Newton's method in terms of computations per iteration.

Comparing Figures 4 and 7, the execution times per iteration of each algorithm between $\psi = 1/32$ and $\psi = 1/8$ are essentially the same, as expected.

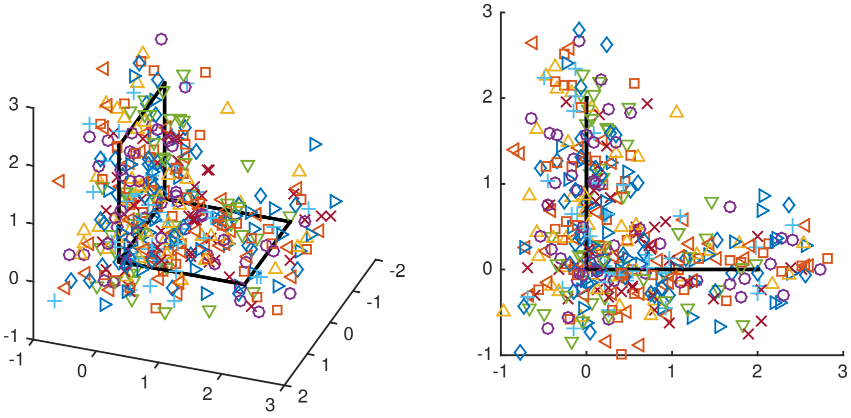

Fig. 5 3-D and 2-D views of simulated sensor data of the simulated MFA with variance $\psi = 1/8$ from $M = 9$ sensor nodes

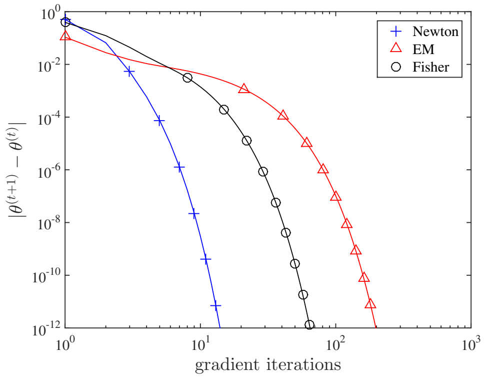

Fig. 6 Change-error vs. gradient iterations from constrained Newton (+), EM (△), and constrained Fisher (○) with uniqueness $\psi = 1/8$

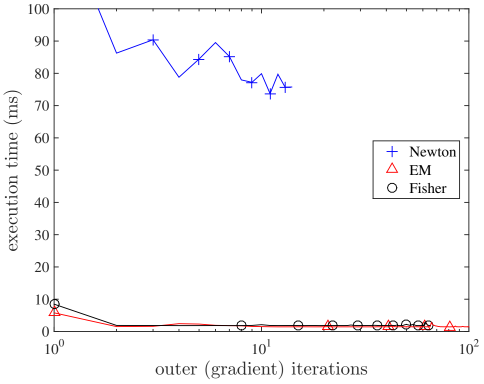

Fig. 7 Execution times per gradient step of constrained Newton (+), EM (△), and constrained Fisher (○) with uniqueness $\psi = 1/8$

4.2 Manifold Learning Example

In this example, we demonstrate that the scoring method benefits from information sharing and integration in model learning. In the previous examples, every node observed, with differing proportions, every factor of the mixture. In this example, the observations depend on the perspective of the sensors relative to the target object. Additionally, the underlying structure is not a finite mixture, but instead a smooth manifold in the form a decaying spiral. Specifically, the decaying spiral is given by

$$\boldsymbol{y} = [(13 - 0.5\omega)\cos\omega, (0.5 - 13)\sin\omega, \omega]^{\mathrm{T}}, \tag{58}$$

where $\omega \in [0, 4\pi)$. Each observation of the spiral is corrupted by additive white Gaussian noise with unit variance. This model was used in previous works to demonstrate the efficacy of learning an MFA model as a surrogate for a nonlinear manifold.[20,21]

The model is set up to illustrate a case in which each node observes a fraction of the entire manifold. Figures 8 and 9 depict this case. Figure 8 shows the samples observed by Node 5, and Figure 9 is an overlay of the observations from all 9 sensors.

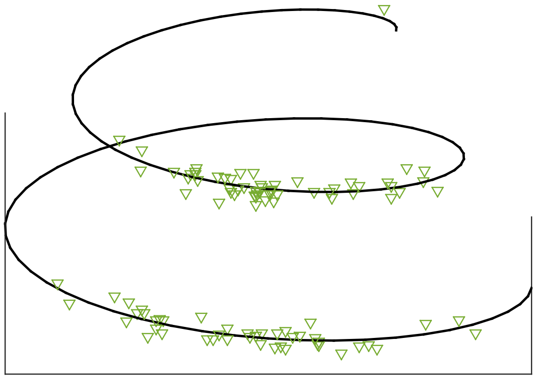

Fig. 8 A realization of noisy samples of the decaying spiral observed by Node 5

We discretize the angles of the spiral so that each sensor observes the spiral according to a discrete probability distribution. The angle distributions as seen by Node 1, 5, and 9 are shown in Fig. 10. As seen in the figure, Node 1 and Node 9 will observe overlapping segments of the spiral with higher probability than Node 5.

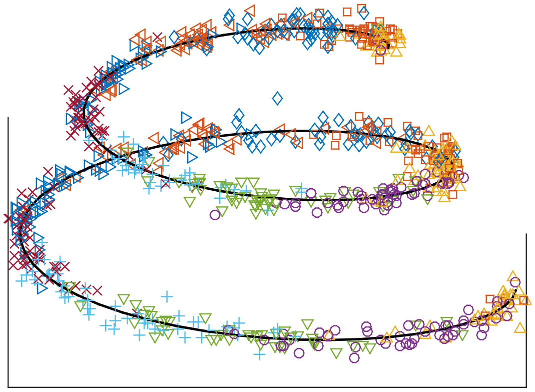

Fig. 9 Samples observed by all sensor nodes. Each marker relates samples to a sensor node

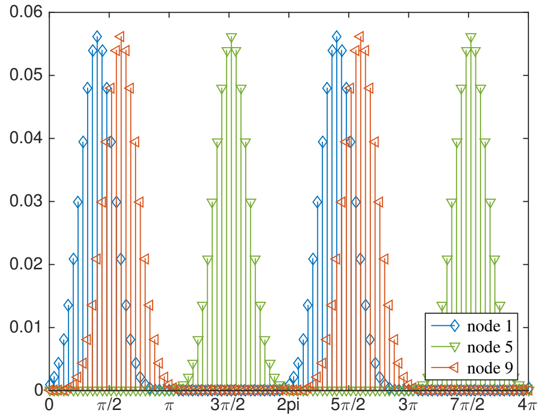

Fig. 10 Distributions of angles of a simulated decaying spiral viewed by sensor nodes. The modes are centered at the sensor's angle relative to the spiral, which is defined over $[0, 4\pi)$, in the spiral's *xy*-plane

For the given set of samples, we compare the results of executing the scoring algorithm at Node 5 with only its data against the centralized scoring method having access to observations from all the nodes. The models are learned with a change-error tolerance of $\epsilon_g = 10^{-6}$. Figures 11 and 12, respectively, display the locally and globally learned tangent vectors of the decaying spiral. By itself, Node 5 learns only a fraction of the overall object. When federated, a central node is able to learn a more rich appearance model of the object despite limited views of the object at each sensor node.

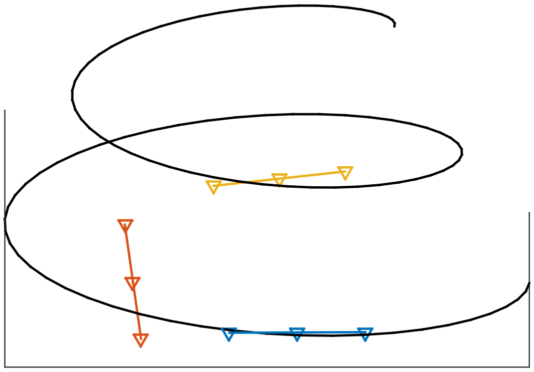

Fig. 11 Estimates of the MFA parameters at Node 5 via Fisher scoring

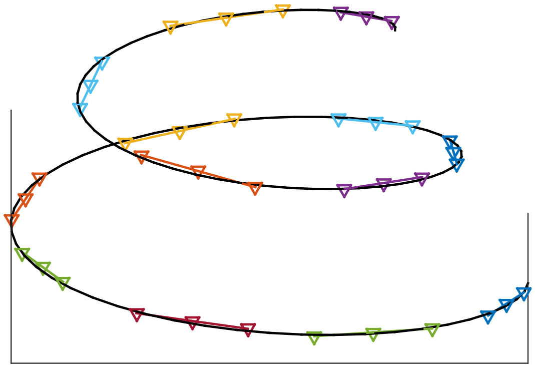

Fig. 12 Estimates of the MFA parameters at a central node via Fisher scoring

5. Conclusions

We derived a constrained Fisher scoring algorithm that exploits block-diagonal and low-rank structures of the expected FIM of the complete data; this results in significantly faster computation per iteration compared to Newton's method. We observed the scoring algorithm converging faster than the EM algorithm at similar per-iteration computation, resulting in speedup factors of 2–10 for the examples considered. Finally, we demonstrated the efficacy of the centralized learning approach for efficiently federating low-rank MFA models across an entire sensor network to provide a global appearance model, even when each sensor has a limited view of the object.

6. References

1. Carin L, Baraniuk RG, Cevher V, Dunson D, Jordan MI, Sapiro G, Wakin MB. Learning low-dimensional signal models. IEEE Signal Processing Magazine. 2011;28(2):39–51.

2. Poor HV. An introduction to signal detection and estimation. 2nd ed. New York (NY): Springer; 1994.

3. Ghahramani Z, Hinton GE. The EM algorithm for mixtures of factor analyzers. Toronto (Canada): University of Toronto; 1996. Report No.: CRG-TR-96-1.

4. Lawley DN, Maxwell AE. Factor analysis as a statistical method. London (UK): Butterworth; 1971.

5. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Royal Statistical Soc Series B. 1977;39:1–38.

6. Redner RA, Walker HF. Mixture densities, maximum likelihood, and the EM algorithm. SIAM review. 1984;26(2):195–239.

7. Jamshidian M, Jennrich RI. Acceleration of the EM algorithm by using quasi-Newton methods. J Royal Statistical Soc. 1997;59(3):569–587.

8. Titterington DM. Recursive parameter estimation using incomplete data. J Royal Statistical Soc Series B (Methodological). 1984:257–267.

9. Xu L, Jordan MI. On convergence properties of the EM algorithm for Gaussian mixtures. Neural Computation. 1996;8(1):129–151.

10. Ramakrishnan N, Ertin E, Moses RL. Gossip-based algorithm for joint signature estimation and node calibration in sensor networks. IEEE J Selected Topics in Signal Processing. 2011;5(4):665–673.

11. Reiersøl O. On the identifiability of parameters in Thurstone's multiple factor analysis. Psychometrika. 1950;15(2):121–149.

12. McLachlan G, Peel D. Mixtures of factor analyzers. In: 7th International Conference on Machine Learning; Burlington (MA): Morgan Kaufmann; 2000. p. 599–606.

13. Anderson TW, Rubin H. Statistical inference in factor analysis. In: Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability; Vol. 5; 1956. p. 111–150.

14. Moore TJ, Sadler BM, Kozick RJ. Maximum-likelihood estimation, the Cramér-Rao bound, and the method of scoring with parameter constraints. IEEE Trans on Signal Processing. 2008;56(3):895–908.

15. Stoica P, Ng BC. On the Cramér-Rao bound under parametric constraints. IEEE Signal Processing Lett. 1998;5(7):177-179.

16. Moore TJ, Kozick RJ, Sadler BM. The constrained Cramér-Rao bound from the perspective of fitting a model. IEEE Signal Processing Letters. 2007;14(8):564–567.

17. Whipps GT, Ertin E, Moses RL. A consensus-based decentralized EM for a mixture of factor analyzers. In: 24th IEEE International Conference on Machine Learning for Signal Processing; 2015, Reims, France.

18. Baek J, McLachlan GJ. Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010;32(7):1298–1309.

19. Bishop CM. Pattern recognition and machine learning. In: Information science and statistics. Berlin (Germany): Spring; 2006.

20. Ueda N, Nakano R, Ghahramani Z, Hinton GE. SMEM algorithm for mixture models. Neural Computation. 2000;12(9):2109–2128.

21. Figueiredo MAT, Jain AK. Unsupervised learning of finite mixture models. IEEE Trans on Pattern Analysis and Machine Intelligence. 2002;24(3):381–396.

Appendix A. Constrained Cramér-Rao Bound

Approved for public release; distribution is unlimited.

Stoica and Ng[1] developed an expression for the constrained Cramér-Rao Bound (CRB) for parametric estimation that does not require the Fisher information matrix (FIM) of the unconstrained problem to be of full rank. The constrained CRB is given by

$$\mathrm{E}\left((\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}\right) \geq U(U^{\mathrm{T}}JU)^{-1}U^{\mathrm{T}}, \tag{A-1}$$

where $\hat{\boldsymbol{\theta}}$ is an unbiased estimate of $\boldsymbol{\theta} \in \mathbb{R}^n$ and J is the FIM of the unconstrained problem. The matrix U is an orthonormal basis for the null space of matrix F defined by

$$F = \frac{\partial \boldsymbol{f}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \in \mathbb{R}^{k\times n} \tag{A-2}$$

where $\boldsymbol{f} \in \mathbb{R}^k$ is a column vector of constraints on $\boldsymbol{\theta}$ such that $f(\boldsymbol{\theta}) = \mathbf{0}$. It is assumed that $\boldsymbol{f}$ satisfies all the conditions in Stoica and Ng[1], such as the number of constrains in $\boldsymbol{f}$ are fewer than the number of parameters in $\boldsymbol{\theta}$ (i.e., $k < n$), and the set of parameters that satisfies the constraints is nonempty. However, as we show next, the basis U need not be orthonormal.

Let the columns of $U \in \mathbb{R}^{n\times n-k}$ form a basis for the null space of F. The matrix defined by

$$P_U = U(U^{\mathrm{T}}U)^{-1}U^{\mathrm{T}} \tag{A-3}$$

projects onto the column space of U. If U is orthonormal, then $P_U = UU^{\mathrm{T}}$ as in Stoica and Ng[1]. Define $\tilde{U} = U(U^{\mathrm{T}}U)^{-1}$, and $W \in \mathbb{R}^{n\times n}$ is an arbitrary matrix. Let $U^{\mathrm{T}}JU$ be positive definite with spectral decomposition given by $U^{\mathrm{T}}JU = QDQ^{\mathrm{T}}$ where Q is an orthogonal matrix and D is diagonal with positive entries along its diagonal. From Eq. 9 in Stoica and Ng[1], we have

$$\begin{aligned}
\mathrm{E}\left((\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^{\mathrm{T}}\right) &\geq WP_U + P_UW^{\mathrm{T}} - WP_UJP_UW^{\mathrm{T}} && \text{(A-4)}\\
&= W\tilde{U}U^{\mathrm{T}} + U(W\tilde{U})^{\mathrm{T}} - (W\tilde{U}Q)D(W\tilde{U}Q)^{\mathrm{T}}\\
&= UQD^{-1}Q^{\mathrm{T}}U^{\mathrm{T}} - (W\tilde{U}Q - UQD^{-1})D(W\tilde{U}Q - UQD^{-1})^{\mathrm{T}}. && \text{(A-5)}
\end{aligned}$$

[1] Stoica P, Ng BC. On the Cramér-Rao bound under parametric constraints. IEEE Signal Processing Letters. 1998;5(7):177-179.

The W that maximizes the right-hand side of Eq. A-4 satisfies

$$W\tilde{U} = UQD^{-1}Q^{\mathrm{T}}. \tag{A-6}$$

By substituting Eq. A-6 into Eq. A-5, we arrive at the lower bound in Eq. A-1 without restricting U to be orthonormal.

Appendix B. Complete Data Fisher Information Matrix for an MFA

In this section, we derive the Fisher information matrix (FIM) for a complete-data model of a mixture of factor analyzers (MFA). We first derive the complete data FIM for a Gaussian mixture model (GMM) and then through a change of variables arrive at the FIM for a MFA.

The complete data log-likelihood of the GMM is given by

$$\ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \log \alpha_{m z_{mi}} \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_{z_{mi}}, \Sigma_{z_{mi}}). \tag{B-1}$$

The complete data FIM of the GMM is defined by

$$J^c_{\boldsymbol{\Phi}} = \mathrm{E}\left(\nabla_{\boldsymbol{\Phi}} \ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z}) \nabla^{\mathrm{T}}_{\boldsymbol{\Phi}} \ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z})\right). \tag{B-2}$$

The gradient of ℓ^c with respect to $\boldsymbol{\Phi} = [\boldsymbol{\mu}_1^{\mathrm{T}}, \boldsymbol{\sigma}_1^{\mathrm{T}}, \ldots, \boldsymbol{\mu}_K^{\mathrm{T}}, \boldsymbol{\sigma}_K^{\mathrm{T}}, \boldsymbol{\alpha}_1^{\mathrm{T}}, \ldots, \boldsymbol{\alpha}_M^{\mathrm{T}}]^{\mathrm{T}}$ is given by

$$\nabla_{\boldsymbol{\Phi}} \ell^c(\boldsymbol{\Phi}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{m=1}^{M} \sum_{i=1}^{N} \begin{bmatrix} \delta(z_{mi} - 1) \nabla_{\boldsymbol{\mu}_1} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_1, \Sigma_1) \\ \delta(z_{mi} - 1) \nabla_{\boldsymbol{\sigma}_1} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_1, \Sigma_1) \\ \vdots \\ \delta(z_{mi} - K) \nabla_{\boldsymbol{\mu}_K} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_K, \Sigma_K) \\ \delta(z_{mi} - K) \nabla_{\boldsymbol{\sigma}_K} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_K, \Sigma_K) \\ \delta(z_{mi} - 1)\delta(m-1)\alpha_{11}^{-1} \\ \vdots \\ \delta(z_{mi} - K)\delta(m-1)\alpha_{1K}^{-1} \\ \vdots \\ \delta(z_{mi} - 1)\delta(m-M)\alpha_{M1}^{-1} \\ \vdots \\ \delta(z_{mi} - K)\delta(m-M)\alpha_{MK}^{-1} \end{bmatrix}, \tag{B-3}$$

where $\delta(a-b)$ is 1 when $a = b$ and 0 otherwise. It is straightforward to show that the gradients in Eq. B-3 are given by

$$\nabla_{\boldsymbol{\mu}_k} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_k, \Sigma_k) = \Sigma_k^{-1}(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k), \tag{B-4}$$

$$\nabla_{\boldsymbol{\sigma}_k} \log \mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_k, \Sigma_k) = \frac{1}{2}\mathrm{vec}\left(\Sigma_k^{-1}(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1} - \Sigma_k^{-1}\right), \tag{B-5}$$

for $k = 1, \ldots, K$. It is clear from Eq. B-3 that the complete data FIM of the GMM is block diagonal, and can therefore be expressed as

$$J_{\mathbf{\Phi}}^{\mathrm{c}} = \text{diag}\left(J_{\boldsymbol{\mu}_1}, J_{\boldsymbol{\sigma}_1}, \ldots, J_{\boldsymbol{\mu}_K}, J_{\boldsymbol{\sigma}_K}, J_{\boldsymbol{\alpha}_1}, \ldots, J_{\boldsymbol{\alpha}_M}\right), \tag{B-6}$$

where

$$\begin{aligned} J_{\boldsymbol{\mu}_k} &= \sum_{m=1}^{M}\sum_{i=1}^{N} \mathrm{E}\left(\delta(z_{mi} - k)\Sigma_k^{-1}(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}\right) \\ &= N\sum_{m=1}^{M} \alpha_{mk}\Sigma_k^{-1}, \qquad \text{(B-7)} \\ J_{\boldsymbol{\alpha}_m} &= \sum_{i=1}^{N} \mathrm{E}\left(\alpha_{mz_{mi}}^{-2}\boldsymbol{e}_{z_{mi}}\boldsymbol{e}_{z_{mi}}^{\mathrm{T}}\right) = N \text{ diag}\left(\boldsymbol{\alpha}_m\right)^{-1}, \qquad \text{(B-8)} \end{aligned}$$

for $k = 1, \ldots, K$ and $m = 1, \ldots, M$. For the term $J_{\boldsymbol{\sigma}_k}$ we use the identity

$$\mathrm{E}\left(\nabla_{\mathbf{\Phi}}\ell^c(\mathbf{\Phi}; \boldsymbol{X}, \boldsymbol{Z})\nabla_{\mathbf{\Phi}}^{\mathrm{T}}\ell^c(\mathbf{\Phi}; \boldsymbol{X}, \boldsymbol{Z})\right) = \mathrm{E}\left(-\nabla_{\mathbf{\Phi}}\nabla_{\mathbf{\Phi}}^{\mathrm{T}}\ell^c(\mathbf{\Phi}; \boldsymbol{X}, \boldsymbol{Z})\right). \tag{B-9}$$

Subsequently, we have for $k = 1, \ldots, K$

$$\begin{aligned} J_{\boldsymbol{\sigma}_k} &= \sum_{m=1}^{M}\sum_{i=1}^{N} \mathrm{E}\left(-\delta(z_{mi} - k)\nabla_{\boldsymbol{\sigma}_k}\nabla_{\boldsymbol{\sigma}_k}^{\mathrm{T}} \log\mathcal{N}(\boldsymbol{x}_{mi}; \boldsymbol{\mu}_k, \Sigma_k)\right) \\ &= \sum_{m=1}^{M}\sum_{i=1}^{N} \mathrm{E}\left(\delta(z_{mi} - k)\frac{1}{2}\left(\Sigma_k^{-1} \otimes \Sigma_k^{-1}(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}\right)\right) \\ &\quad + \sum_{m=1}^{M}\sum_{i=1}^{N} \mathrm{E}\left(\delta(z_{mi} - k)\frac{1}{2}\left(\Sigma_k^{-1}(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)(\boldsymbol{x}_{mi} - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1} \otimes \Sigma_k^{-1}\right)\right) \\ &\quad - \sum_{m=1}^{M}\sum_{i=1}^{N} \mathrm{E}\left(\delta(z_{mi} - k)\frac{1}{2}\left(\Sigma_k^{-1} \otimes \Sigma_k^{-1}\right)\right) \\ &= \frac{N}{2}\sum_{m=1}^{M} \alpha_{mk}\left(\Sigma_k^{-1} \otimes \Sigma_k^{-1}\right). \qquad \text{(B-10)} \end{aligned}$$

We have established the block-diagonal form of the complete data FIM of the GMM. Now we relate the FIM of the MFA to the FIM of the GMM through a change of variables. The factor means and mixing proportions remain unchanged between the GMM and MFA. The change is limited to the parameterization of each factor's covariance through the structure given in Eq. 3. The Jacobian matrix for the transformation is given by

$$G_{\boldsymbol{\theta}} = \frac{\partial \boldsymbol{\Phi}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = \begin{bmatrix} \mathrm{I}_p & \mathbf{0} & & & & \\ \mathbf{0} & G_{\boldsymbol{\sigma}_1} & & & & \\ & & \ddots & & & \\ & & & \mathrm{I}_p & \mathbf{0} & \\ & & & \mathbf{0} & G_{\boldsymbol{\sigma}_K} & \\ & & & & & \mathrm{I}_{MK} \end{bmatrix}, \tag{B-11}$$

where $G_{\boldsymbol{\sigma}_k} \in \mathbb{R}^{p^2 \times p(r+1)}$ is defined in Eq. 33. The complete data FIM of the MFA is then given by

$$J^c_{\boldsymbol{\theta}} = \mathrm{diag}\left(J_{\boldsymbol{\mu}_1}, (G^{\mathrm{T}}_{\boldsymbol{\sigma}_1} J_{\boldsymbol{\sigma}_1} G_{\boldsymbol{\sigma}_1}), \ldots, J_{\boldsymbol{\mu}_K}, (G^{\mathrm{T}}_{\boldsymbol{\sigma}_K} J_{\boldsymbol{\sigma}_K} G_{\boldsymbol{\sigma}_K}), J_{\boldsymbol{\alpha}_1}, \ldots, J_{\boldsymbol{\alpha}_M} \right). \tag{B-12}$$

List of Symbols, Abbreviations, and Acronyms

Terms:

EM	expectation-maximization
FIM	Fisher information matrix
GMM	Gaussian mixture model
MFA	mixture of factor analyzers
ML	maximum likelihood

Mathematical symbols:

α_{mk}	mixing proportion of the k^{th} factor observed by the m^{th} sensor node
Λ_k	loading matrix of the k^{th} factor
$\boldsymbol{\mu}_k$	mean of the k^{th} factor
$\mathbf{\Phi}$	vector of unknown parameters of the GMM model
ψ_k	uniqueness of the k^{th} factor
Σ_k	covariance of the k^{th} factor
$\boldsymbol{\theta}$	vector of unknown parameters of the MFA model
$\boldsymbol{x}_{mi}$	i^{th} observation from the m^{th} sensor node

Mathematical operators:

diag $(A_1, A_2, \dots, A_n)$	a diagonal matrix with $A_1, A_2, \dots, A_n$ along the diagonal where each A_i may be a square matrix with differing sizes across $i = 1, 2, \dots, n$
E()	the expected value of a random quantity
$\max(\ )$	the maximum of an otherwise variable quantity
$\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$	the gradient of function f with respect to $\boldsymbol{\theta}$
$(A \otimes B)$	the Kronecker product of matrix A and B
vec (A)	the vectorization of matrix A. If $A = [\boldsymbol{a}_1, \boldsymbol{a}_2, \dots, \boldsymbol{a}_n]$, then vec $(A) = [\boldsymbol{a}_1^{\mathrm{T}}, \boldsymbol{a}_2^{\mathrm{T}}, \dots, \boldsymbol{a}_n^{\mathrm{T}}]^{\mathrm{T}}$

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
4 (PDF)	DIR USARL RDRL SES A J GEORGE L KAPLAN B RIGGAN N SROUR
2 (PDF)	OHIO STATE UNIV E ERTIN R MOSES

INTENTIONALLY LEFT BLANK.